What Can We Learn from Collective Human Opinions on Natural Language Inference Data?

Yixin Nie　　**Xiang Zhou**　　**Mohit Bansal**
Department of Computer Science
University of North Carolina at Chapel Hill
{yixin1, xzh, mbansal}@cs.unc.edu

Abstract

Despite the subjective nature of many NLP tasks, most NLU evaluations have focused on using the majority label with presumably high agreement as the ground truth. Less attention has been paid to the distribution of human opinions. We collect **ChaosNLI**, a dataset with a total of 464,500 annotations to study **C**ollective **H**um**A**n **O**pinion**S** in oft-used **NLI** evaluation sets. This dataset is created by collecting 100 annotations per example for 3,113 examples in SNLI and MNLI and 1,532 examples in αNLI. Analysis reveals that: (1) high human disagreement exists in a noticeable amount of examples in these datasets; (2) the state-of-the-art models lack the ability to recover the distribution over human labels; (3) models achieve near-perfect accuracy on the subset of data with a high level of human agreement, whereas they can barely beat a random guess on the data with low levels of human agreement, which compose most of the common errors made by state-of-the-art models on the evaluation sets. This questions the validity of improving model performance on old metrics for the low-agreement part of evaluation datasets. Hence, we argue for a detailed examination of human agreement in future data collection efforts, and evaluating model outputs against the distribution over collective human opinions.[1]

1 Introduction

Natural Language Understanding (NLU) evaluation plays a key role in benchmarking progress in natural language processing (NLP) research. With the recent advance in language representative learning (Devlin et al., 2019), results on previous benchmarks have rapidly saturated. This leads to an explosion of difficult, diverse proposals of tasks/datasets for NLU evaluation, including Natural Language Inference (e.g., SNLI, MNLI and ANLI) (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020), Grounded Commonsense Inference (Zellers et al., 2018), Commonsense QA (Talmor et al., 2019), Social Interactions Reasoning (Sap et al., 2019), Abductive Commonsense Reasoning (αNLI) (Bhagavatula et al., 2020), etc.

One common practice followed by most of these recent works is to simplify the evaluation of various reasoning abilities as a classification task. This is analogous to asking objective questions to a human in educational testing. This simplification not only facilitates the data annotation but also gives interpretable evaluation results, based on which behaviors of the models are studied and then weaknesses are diagnosed (Sanchez et al., 2018).

Despite the straightforwardness of this formalization, one assumption behind most prior benchmark data sourcing is that there exists a single prescriptive ground truth label for each example. The assumption might be true in human educational settings where prescriptivism is preferred over descriptivism because the goal is to test humans with well-defined knowledge or norms (Trask, 1999). However, it is not true for many NLP tasks due to their pragmatic nature where the meaning of the same sentence might differ depending on the context or background knowledge.

Specifically for the NLI task, Manning (2006) advocate that annotation tasks should be "natural" for untrained annotators, and the role of NLP should be to model the inferences that humans make in practical settings. Previous work (Pavlick and Kwiatkowski, 2019) that uses a graded labeling schema on NLI, showed that there are inherent disagreements in inference tasks. All these discussions challenge the commonly used majority "gold-label" practice in most prior data collections and evaluations.

Intuitively, such disagreements among humans

---

[1]The ChaosNLI dataset and experimental scripts are available at https://github.com/easonnie/ChaosNLI

should be allowed because different annotators might have different subjective views of the world and might think differently when they encounter the same reasoning task. Thus, from a descriptive perspective, evaluating the capacity of NLP models in predicting not only individual human opinions or the majority human opinion, but also the overall distribution over human judgments provides a more representative comparison between model capabilities and 'collective' human intelligence.

Therefore, we collect **ChaosNLI**, a large set of **C**ollective **H**um**A**n **O**pinion**S** for examples in several existing (English) NLI datasets, and comprehensively examine the factor of human agreement (measured by the entropy of the distribution over human annotations) on the state-of-the-art model performances. Specifically, our contributions are:

- We collect additional 100 annotations for over 4k examples in SNLI, MNLI-matched, and αNLI (a total of 464,500 annotations) and show that when the number of annotations is significantly increased: (1) a number of original majority labels fail to present the prevailing human opinion (in 10%, 20%, and 31% of the data we collected for αNLI, SNLI, and MNLI-matched, respectively), and (2) large human disagreements exist and persist in a noticeable amount of examples.

- We compare several state-of-the-art model[2] outputs with the distribution of human judgements and show that: (1) the models lack the ability to capture the distribution of human opinions[3]; (2) such ability differs from the ability to perform well on the old accuracy metric; (3) models' performance on the subset with high levels of human agreement is substantially better than their performance on the subset with low levels of human agreement (almost close to solved versus random guess, respectively) and shared mistakes by the state-of-the-art models are made on examples with large human disagreements.

- We argue for evaluating the models' ability to predict the distribution of human opinions and discuss the merit of such evaluation with respect to NLU evaluation and model calibration. We also give design guidance on crowd-sourcing such collective annotations to facilitate future studies on relevant pragmatic tasks.

---

[2]We test models including BERT, RoBERTa, XLNET, ALBERT, DistilBERT, and BART.

[3]We measure the Jensen-Shannon Distance (JSD) and the Kullback–Leibler (KL) divergence between model softmax outputs and the estimated distribution over human annotations.

The **ChaosNLI** dataset and experimental scripts are available at `https://github.com/easonnie/ChaosNLI`

2 Related Work

Uncertainty of Annotations. Past discussions of human disagreement on semantic annotation tasks were mostly focused on the uncertainty of individual annotators and the noisiness of the data collection process. These tasks include word sense disambiguation (Erk and McCarthy, 2009; Jurgens, 2013), coreference (Versley, 2008), frame corpus collection (Dumitrache et al., 2019), anaphora resolution (Poesio and Artstein, 2005; Poesio et al., 2019), entity linking (Reidsma and op den Akker, 2008), tagging and parsing (Plank et al., 2014; Alonso et al., 2015), and veridicality (De Marneffe et al., 2012; Karttunen et al., 2014). These works focused on studying the ambiguity of annotations, how the design of the annotation setup might affect the inter-annotator-agreement, and how to make the annotations reliable. However, we consider the disagreements and subjectivity to be an intrinsic property of the populations. Our work discusses the disagreements among a large group of individuals, and further examines the relation between the annotation disagreement and the model performance.

Disagreements in NLI Annotations. Our work is significantly inspired by previous work that reveals the "inherent disagreements in human textual inference" (Pavlick and Kwiatkowski, 2019). It employed 50 independent annotators for a "graded" textual inference task, yielding a total of roughly 19,840 annotations, and validates that disagreements among the annotations are reproducible signals. In particular, in their work, the labeling schema is modified from 3-way categorical NLI to a graded one, whereas our study keeps the original 3-way labeling schema to facilitate a direct comparison between old labels and new labels, and focuses more on giving an in-depth analysis regarding the relation between the level of disagreements among humans and the state-of-the-art model performance.

Graded Labeling Schema. Some previous work attempts to address the issues with human disagreements by modifying or re-defining the evaluation task with a more fine-grained ordinal or even real-value labeling schema rather than categorical labeling schema (Zhang et al., 2017; Pavlick and

Kwiatkowski, 2019; Chen et al., 2019) to reduce the issues of ambiguity. Our work is independent and complementary to those by providing analysis on general language understanding from a collective distribution perspective.

3 Data Collection

Our goal is to gather annotations from multiple annotators to estimate the distribution over human opinions. Section 3.1 and 3.2 state some details of the collection. More importantly, Section 3.3 explains the challenges of such data collection and how our designs ensure data quality.

3.1 Dataset and Task

ChaosNLI provides 100 annotations for each example in three sets of existing NLI-related datasets. The first two sets are a subset of the SNLI development set and a subset of MNLI-matched development set, in which the examples satisfy the requirement that their majority label agrees with only three out of five individual labels collected by the original work (Bowman et al., 2015; Williams et al., 2018).[4] The third set is the entire αNLI development set introduced in Bhagavatula et al. (2020). To simplify the terminology, we denote SNLI, MNLI-m and αNLI portion of the ChaosNLI as ChaosNLI-S, ChaosNLI-M, and ChaosNLI-α, respectively.

3.2 Annotation Interface

To collect multiple labels for each example, we employed crowdsourced workers from Mechanical Turker with qualifications. The annotation interface is implemented using the ParlAI[5] (Miller et al., 2017) framework. The collection is embodied in a multi-round interactive environment where at each round annotators are instructed to do one single multi-choice selection. This reduces the annotators' mental load and helps them focus on the human intelligence tasks (HITs). The compressed versions of instructions are shown in Figure 1. Screenshots of Turker interfaces are attached in Appendix A.

3.3 Quality Control

Collecting the "soft-label" for examples based on plausible human opinions is difficult because we

---

[4]All the examples in SNLI and MNLI development and test set come with 5 labels and the ground truth labels are defined by majority label in all previous studies. Here, we intentionally choose to label examples with a low level of human agreement in SNLI and MNLI to highlight the factor of human disagreement. Both datasets are in English.

[5]https://parl.ai/

Natural Language Inference (NLI)

> Given a context, a statement can be either:
> - Definitely correct (Entailment); or
> - Definitely incorrect (Contradiction); or
> - Neither (Neutral).
>
> Your goal is to choose the correct category for a given pair of context and statement.
> An automatic detector will estimate your annotation accuracy on this task. If your estimated accuracy is **too low**, you might be **disqualified**.
> If you feel **uncertain** about some examples, just choose the best category you believe the statement should be in.
>
> *Examples:*
> ***Context:*** *A guitarist is playing in a band.*
> ***Statement:*** *Some people are performing.*
> ***Answer:*** *The statement is* ***definitely correct***.

Abductive Natural Language Inference (αNLI)

> Given two observations (**O-Beginning** and **O-Ending**), and two hypotheses (**H1** and **H2**), your goal is to choose one of the hypotheses that is more likely to cause **O-Beginning** to turn into **O-Ending**.
> An automatic detector will estimate your annotation accuracy on this task. If your estimated accuracy is **too low**, you might be **disqualified**.
> If you feel **uncertain** about some examples, just choose the best category you believe the statement should be in.
>
> *Examples:*
> ***O-Beginning:*** *Jenny cleaned her house and went to work, leaving the window just a crack open.*
> ***H1:*** *A thief broke into the house by pulling open the window.*
> ***H2:*** *Her husband went home and close the window.*
> ***O-Ending***: *When Jenny returned home she saw that her house was a mess.*
> ***Answer: H1***.

Figure 1: Mechanical Turker instructions (compressed) for NLI and αNLI.

need to enforce that each annotator will genuinely try their best on the work to avoid errors caused by carelessness. We can not denoise the data by collecting more annotations and aggregating them with majority voting, nor can we use the inter-annotator agreement to measure data quality.

To this end, we select a set of examples , which exhibit high human agreement for a single label, to rigorously test and track the performance of each annotator. We call them ***the set of unanimous examples***. To obtain such set, we sampled examples from SNLI, MNLI, and αNLI training set, then crowdsourced 50 annotations for each of them, and finally selected those whose human agreement is indeed high (majority>95%). Throughout the collection process, we employ the following three mechanisms to ensure label quality:

On-boarding test. Every annotator needs to pass an on-boarding test before they can work on any real example. The test includes five easy examples pre-selected from ***the set of unanimous examples***. If they fail to give the correct selection for any of them, they will be prevented from working on any example. The mechanism tests whether the annotator understands the task.

Training phase. After passing the on-boarding test, each annotator will be given 10 to 12 examples from *the set of unanimous examples* to be further annotated. For each example, if an annotator gives a label that is different from the pre-collected legitimate label, the annotator will be prompted with the correct label and told to keep concentrating on the task. If the accuracy of an annotator on training examples is below 75%, the annotator will be disallowed to proceed. This training mechanism further helps the annotators get familiar with the task.

Performance tracking. After the training phase, annotators will be given real examples. For each example to be annotated, there will be 10% chance that the example is sampled from *the set of unanimous examples*. Again, for such examples, if an annotator gives a label that is different from the pre-collected legitimate label, the annotator will be prompted with the correct label and told to keep concentrating on the task. If the accuracy of an annotator on those examples is below 75% or if the annotator gives four consecutive incorrect labels, the annotator will be blocked. This mechanism tracks the performance of each annotator and guarantees that each annotator is capable and focused when working on any examples.

Table 1 shows that on-boarding test filters more than half of the turkers. Figure 2 shows that the average accuracy of a single Turker on *the set of unanimous examples* improves as the annotators have completed more examples and converges at around 92%.[6] The observations indicate that our filtering mechanisms are rigorous and help improving and keeping annotator concentrate during the collection task. The design gives guidelines for future work on how to ensure data quality where normal inter-annotator-agreement measures are not applicable.

3.4 Other Details

The entire collection takes about one month to complete over 464K annotations. The mean/median time a turker spent on each example ranges from 10 to 20 seconds as shown in Table 1 (and we pay up to $0.5 on average per HIT of ten examples). We observe high variance in the time/example across turkers (including over-estimation due to breaks), hence the median estimate is more reliable. We

---

[6]This is comparable to the accuracy of a majority voting over 5 aggregated annotations in previous work (Nangia and Bowman, 2019).
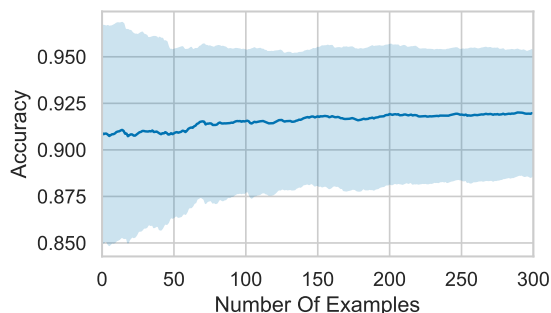

Figure 2: The accuracy range of the annotators on the NLI training and hidden *unanimous examples* as they annotated their first 300 examples.

Data	QFR (%)	FR (%)	#Turkers	Time (sec)
ChaosNLI-α	7.4	1.3	1,903	18.7 / 12.7
ChaosNLI-S	39.9	14.1	1,639	15.9 / 10.1
ChaosNLI-M	39.9	14.1	1,744	21.2 / 13.3

Table 1: MTurk statistics on the three datasets. 'QFR' or Qualification Fail Rate refers to the failure rate of the onboarding qualification test. 'FR' or Filter Rate refers to the ratio of Turkers who got blocked (during training phase and performance tracking described in Section 3.3) because their performance on the *unanimous examples* set are too low. SNLI and MNLI-m shared the same onboarding test and the same *unanimous examples* set, therefore their numbers are the same. The '#Turkers' column denotes the final set of filtered turkers that contributed to the released annotations. The last column 'Time' refers to the mean / median time spent by Turkers per example in seconds.

had a large set of qualified turkers for our final annotations. The total time of one month is largely attributed to the rigorous quality control mechanism via careful on-boarding qualification tests and quality monitoring.

4 Analysis of Human Judgements

Statistics. We collected 100 new annotations for each example in the three sets described in Section 3.1. Table 2 shows the total number of examples in the three sets and the percentage of cases where the new majority label is different from the old majority label (based on 5 annotations for SNLI and MNLI and 1 annotation for αNLI in their original dataset, respectively). Since we only collected labels for subsets of SNLI and MNLI-m, we also include the size of the original SNLI and MNLI-m development sets and the change-of-label ratio with respect to the original sets. The findings suggest that the old labels fail to present the genuine majority labels among humans for a noticeable amount
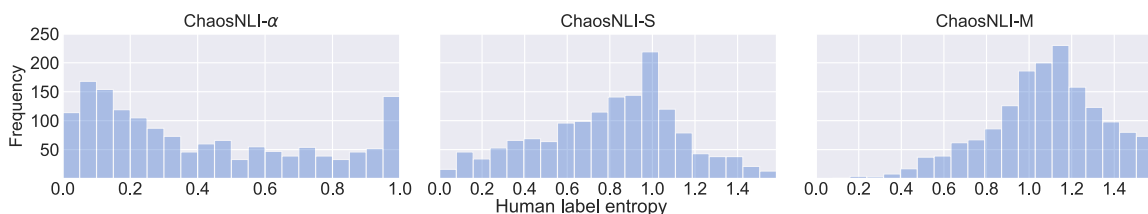
Figure 3: Histogram of entropy of estimated distribution over human annotations on ChaosNLI-α, ChaosNLI-S, ChaosNLI-M.

Data	# Examples	Change rate (%)
ChaosNLI-α	1,532	10.64
ChaosNLI-S	1,514 (10k)	24.97 (3.78)
ChaosNLI-M	1,599 (10k)	31.77 (5.08)

Table 2: Data Statistics. '# Examples' refers to the total number of examples. 'Change rate' refers to the percentage that the old majority label is different from the new majority label. The number in the parentheses shows the size of the entire original SNLI and MNLI-m development set and the percentage of label changes with respect to the entire set.

(10%, 25%, and 30% for ChaosNLI-α, ChaosNLI-S, and ChaosNLI-M, respectively) of the data. The label statistics for individual datasets can be found in Appendix D.

Examples. Table 3 and Table 4 show some collected NLI examples that either have low levels of human agreements or have different majority labels as opposed to the old ground truth labels. We can see that the resultant labels we collected not only provide more fine-grained human judgements but also give a new majority label that is better at presenting the prevailing human opinion. Moreover, there indeed exist different but plausible interpretations for the examples that are of low-level of human agreements and the discrepancy is not just noise but presents the distribution over human judgements with "higher resolutions". This is consistent with the finding in Pavlick and Kwiatkowski (2019).

Entropy Distribution. To further investigate the human uncertainty in our collected labels, we show the histogram of the entropy of label distribution for ChaosNLI-α, ChaosNLI-S and ChaosNLI-M in Figure 3. The label distribution is approximated by the 100 collected annotations. The entropy is calculated with $\mathbf{H}(\mathbf{p}) = -\sum_{i \in \mathcal{C}} p_i \log(p_i)$ and $p_i = \frac{n_i}{\sum_{j \in \mathcal{C}} n_j}$, where $\mathcal{C}$ is the label category set and n_i is the number of labels for category i. The

entropy value gives a measure for the level of uncertainty or agreement among human judgements, where high entropy suggests low level of agreement and vice versa.

The histogram for the ChaosNLI-α shows a distribution that is similar to a U-Shaped distribution. This indicates that naturally occurring examples in ChaosNLI-α are either highly certain or uncertain among human judgements. In ChaosNLI-S and ChaosNLI-M, the distribution shows only one apparent peak; and the distribution for ChaosNLI-M is slightly skewed towards higher entropy direction. As described in Section 3.1, ChaosNLI-S and ChaosNLI-M are subsets of SNLI and MNLI-m development that are of low-level of human agreements, it could be expected that the majority of naturally occurring SNLI and MNLI data would also have low entropy, which will form another peak around the beginning of the x-axis resulting a U-like shape similar to ChaosNLI-α.[7]

5 Analysis of Model Predictions

In Section 4, we discussed the statistics and some examples for the new annotations. The observation naturally raises two questions regarding the development of NLP models: (1) whether the state-of-the-art models are able to capture this distribution over human opinions; and (2) how the level of human agreements will affect the performance of the models. Hence, we investigate these questions in this section. Section 5.1 and 5.2 state our experimental choices. Section 5.3 discusses the results regarding the extent to which the softmax distributions produced by state-of-the-art models trained on the dataset reflects similar distributions over human annotations. Section 5.4 demonstrate the surprising influence of human agreements on the model performances.

---

[7]In our pilot study, we collected 50 labels for 100 examples of SNLI where all five original annotators agreed with each other, the average entropy of those is 0.31. The average entropy of examples on ChaosNLI-S is 0.80.

Context	Hypothesis	Old Labels	New Labels majority and individual labels	Source	Type
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** E E E N N	**Entailment** E[(50)] N[(50)]	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** E E N N E	**Neutral** N[(93)] E[(7)]	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** C N N C C	**Contradiction** C[(56)] N[(44)]	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** N E N N E	**Entailment** E[(94)] N[(5)] C[(1)]	MNLI	Majority changed

Table 3: Examples from ChaosNLI-S and ChaosNLI-M development set. 'Old Labels' is the 5 label annotations from original dataset. 'New Labels' refers to the newly collected 100 label annotations. Superscript indicates the frequency of the label.

Observation-1	Sadie was on a huge hike.
Observation-2	Luckily she pushed herself and managed to reach the peak.
Hypothesis-1	Sadie almost gave down mid way.
Hypothesis-2	Sadie wanted to go to the top.
Old Label	Hypothesis-2
New Labels	**Hypothesis-1[(58)]** Hypothesis-2[(42)]
Observation-1	Uncle Jock couldn't believe he was rich.
Observation-2	Jock lived the good life for a whole year, until he was poor again.
Hypothesis-1	He went to town and spent on extravagant things.
Hypothesis-2	Jock poorly managed his finances.
Old Label	Hypothesis-1
New Labels	Hypothesis-1[(48)] **Hypothesis-2[(52)]**

Table 4: Examples from the collected ChaosNLI-α development set. The task asks which of the two hypothesis is more likely to cause **Observation-1** to turn into **Observation-2**. Superscript indicates the frequency of the label. Majority labels were marked in bold.

5.1 Models and Setup

Following the pretraining-then-finetuning trend, we focus our experiments on large-scale language pretraining models. We studied BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019) since they are considered to be the state-of-the-art models for learning textual representations and have been used for a variety of downstream tasks. We experimented on both the base and the large versions of these models, in order to analyze the parameter size factor. Additionally, we include BART (Lewis et al., 2020), ALBERT (Lan et al., 2019), and DistilBERT (Sanh et al., 2019) in the experiments. ALBERT is designed to reduce parameters of BERT by cross-layer parameter sharing and decomposing embedding. DistilBERT aims to compress BERT with knowledge distillation. BART is a denoising autoencoder for pretraining seq-to-seq models.

For NLI, we trained the models on a combined training set of SNLI and MNLI which contains over 900k NLI pairs. We used the best hyper-parameters chosen by their original authors. For αNLI, we trained the models on αNLI training set (169,654

examples). The hyper-parameters for αNLI were tuned with results on αNLI development set. Details of the hyper-parameters are in Appendix B.

5.2 Evaluation and Metrics

As formulated in Equation 4, we used the 100 collected annotations for each example to approximate the human label distributions for each example. In order to examine to what extent the current models are capable of capturing the collective human opinions, we compared the human label distributions with the softmax outputs of the neural networks following Pavlick and Kwiatkowski (2019).

We used Jensen-Shannon Distance (JSD) as the primary measure of the distance between the softmax multinomial distribution of the models and the distributions over human labels because JSD is a metric function based on a mathematical definition (Endres and Schindelin, 2003). It's symmetric and bounded with the range [0, 1], whereas the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951; Kullback, 1997) does not have these two properties. We also used KL as a complementary measure. The two metrics are calculated as:

$$\text{KL}\left(\mathbf{p}\|\mathbf{q}\right) = \sum_{i\in\mathcal{C}} p_i \log\left(\frac{p_i}{q_i}\right) \quad (1)$$

$$\text{JSD}\left(\mathbf{p}\|\mathbf{q}\right) = \sqrt{\frac{1}{2}(\text{KL}\left(\mathbf{p}\|\mathbf{m}\right) + \text{KL}\left(\mathbf{q}\|\mathbf{m}\right))} \quad (2)$$

where $\mathbf{p}$ is the estimated human distribution, $\mathbf{q}$ is model softmax outputs, and $\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q})$.

5.3 Main Results

Table 5 reports the main results regarding the distance between model softmax distribution and estimated human label distribution. In addition to the models, we also show the results for the chance baseline (the first row) and the results for estimated

Model	ChaosNLI-α			ChaosNLI-S			ChaosNLI-M		
	JSD↓	KL↓	Acc.↑ (old/new)	JSD↓	KL↓	Acc.↑ (old/new)	JSD↓	KL↓	Acc.↑ (old/new)
Chance	0.3205	0.406	0.5098/0.5052	0.383	0.5457	0.4472/0.5370	0.3023	0.3559	0.4509/0.4634
BERT-b	0.3209	3.7981	0.6527/0.6534	0.2345	0.481	0.7008/0.7292	**0.3055**	<u>0.7204</u>	0.5991/0.5591
XLNet-b	0.2678	<u>1.0209</u>	0.6743/0.6867	0.2331	0.5121	0.7114/0.7365	<u>0.3069</u>	0.7927	0.6373/0.5891
RoBERTa-b	0.2394	**0.8272**	0.7154/0.7396	0.2294	0.5045	0.7272/0.7536	0.3073	0.7807	0.6391/0.5922
BERT-l	0.3055	3.7996	0.6802/0.6821	0.23	0.5017	0.7266/0.7384	0.3152	0.8449	0.6123/0.5691
XLNet-l	0.2282	1.8166	0.814/0.8133	0.2259	0.5054	<u>0.7431</u>/0.7807	0.3116	0.8818	**0.6742**/<u>0.6185</u>
RoBERTa-l	**0.2128**	1.3898	**0.8531**/<u>0.8368</u>	<u>0.221</u>	0.4937	**0.749/0.7867**	0.3112	0.8701	**0.6742/0.6354**
BART	0.2215	1.5794	0.8185/0.814	**0.2203**	<u>0.4714</u>	0.7424/<u>0.7827</u>	0.3165	0.8845	<u>0.6635</u>/0.5922
ALBERT	<u>0.2208</u>	2.9598	<u>0.8440</u>/**0.8473**	0.235	0.5342	0.7153/0.7814	0.3159	0.862	0.6485/0.5897
DistilBert	0.3101	1.0345	0.592/0.607	0.2439	**0.4682**	0.6711/0.7021	0.3133	**0.6652**	0.5472/0.5103
Est. Human	0.0421	0.0373	0.885/0.97	0.0614	0.0411	0.775/0.94	0.0695	0.0381	0.66/0.86

Table 5: Model Performances for JSD, KL, and Accuracy on majority label. ↓ indicates smaller value is better. ↑ indicates larger value is better. For each column, the best values are in bold and the second best values are underlined. "-b" and "-l" in the Model column denote "-base" and "-large", respectively.

human performance (the last row). The chance baseline gives each label equal probability when calculating the JSD and KL measures. The accuracy of the chance baseline directly shows the proportion of the examples with the majority label in a specific evaluation set. To estimate the human performance, we employed a new set of annotators to collect another 100 labels for a set of randomly sampled 200 examples on ChaosNLI-α, ChaosNLI-S and ChaosNLI-M, respectively. For a better estimation of 'collective' human performance, we ensure that the new set of annotators employed for estimating human performance is disjoint from the set of annotators employed for the normal label collection.[8] In what follows, we discuss the results.

Significant difference exists between model outputs and human opinions. The most salient information we can get is that there are large gaps between model outputs and human opinions. To be specific, the estimated collective human performance gives JSD and KL scores far below 0.1 on all three sets. However, the best JSD achieved by the models is larger than 0.2 and the best KL achieved by the models barely goes below 0.5 across the table. The finding can be somewhat foreseeable since none of the models are designed to capture collective human opinions and suggests room for improvement.

Even chance baseline is hard to beat. What is more surprising is that a number of these state-of-the-art models can barely outperform and sometimes even perform worse than the chance baseline w.r.t. JSD and KL scores. On ChaosNLI-M, all the models yield similar JSD scores to the chance baseline and are beaten by it on KL. On ChaosNLI-α, BERT-base performs worse than the chance baseline on JSD and the scores of KL by all the models are way higher than that of the chance baseline. This hints that capturing human label distribution is a common *blind spot* for many models.

There is no apparent correlation between the accuracy and the two divergence scores. On both ChaosNLI-S and ChaosNLI-M, DistilBERT gives the best KL scores despite the fact that it obtains the lowest accuracy on the majority label. BERT-base gives the best JSD while having the second lowest accuracy on ChaosNLI-M. RoBERTa-large gives the highest accuracy on ChaosNLI-S and ChaosNLI-M, and the second highest accuracy on ChaosNLI-α but it only obtains the lowest JSD on ChaosNLI-α. The best JSD score on ChaosNLI-α is achieved by BART but it fails to give the best accuracy. This hints that the ability required to model the distribution of human labels differs from that required to predict the majority label and perform well on the accuracy metric.

Large models are not always better. Direct comparison between base and large models for BERT, XLNet, and RoBERTa reveals that large models cannot beat base models on ChaosNLI-α and ChaosNLI-M on KL scores. Moreover, on

---

[8]The estimation of collective human performance can also be viewed as calculating the JSD and KL between two disjoint sets of 100 human opinions.
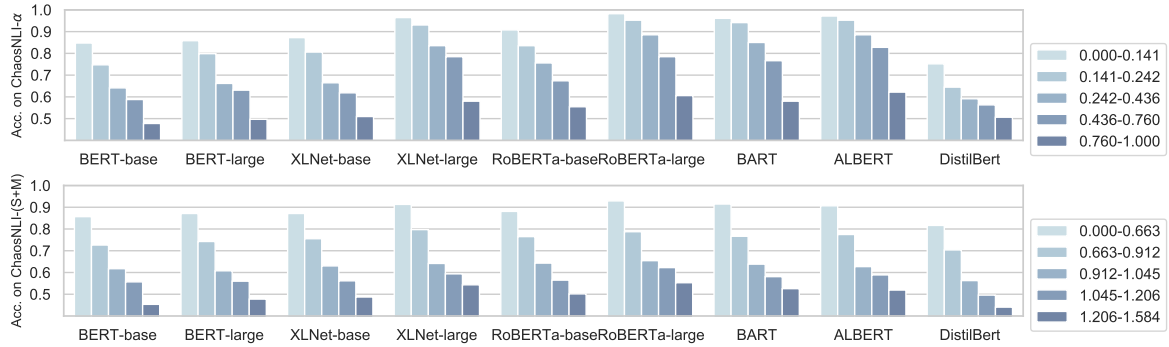
Figure 4: Accuracy on different bins of data points whose entropy values are within specific quantile ranges.
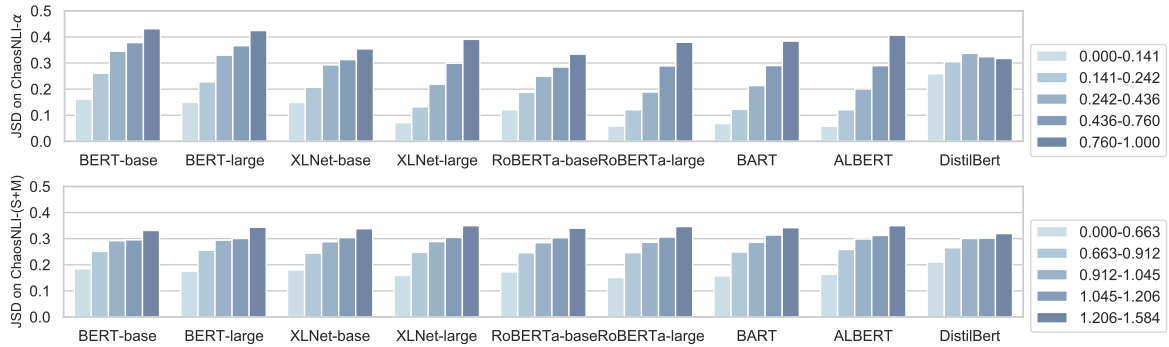

Figure 5: JSD on different bins of data points whose entropy values are within specific quantile ranges.

ChaosNLI-M, all the large models give higher JSD scores than the base models. However, all the large models achieve higher accuracy than their base model counterparts on all three evaluation sets. This observation suggests that modeling the collective human opinions might require more thoughtful designs instead of merely increasing model parameter size.

5.4 The Effect of Agreement

To study how human agreements will influence the model performance, we compute the entropy of the human label distribution (by Equation 4) for each data point. Then, we partition ChaosNLI-α and the union of ChaosNLI-S and ChaosNLI-M using their respective entropy quantiles as the cut points. This results in several bins with roughly equal numbers of data points whose entropy lies in a specific range. Figure 4 and 5 shows the accuracy and the JSD of the models on different bins.[9] We observe that:

- Across the board, *there are consistent correla-*

[9]Model JSD performances are similar to the accuracy performances where all the models obtain worse results at the bins with higher entropy range. One exception is the JSD of DistilBert on ChaosNLI-α. This might due to the fact that DistilBert is highly uncertain in its prediction and tend to give even distribution for each label yielding similar results to the chance baseline.

tions between the level of human agreements and the accuracy of the model. This correlation is positive, meaning that all models perform well on examples with a high level of human agreements while struggle with examples having a low level of human agreements. Similar trends also exists in JSD.

- Accuracy downgrades dramatically (from 0.9 to 0.5) as the level of human agreements decrease.

- The model barely outperforms and sometimes even under-performs the chance baseline on bins with the lowest level of human agreements. For both αNLI and NLI, the accuracy of most models on the bin with the lowest level of human agreements does not surpass 60%.

These results reveal that most of the data (which often compose the majority of the evaluation set) with a high level of human agreement have been solved by state-of-the-art models, and most of the common errors on popular benchmarks (like αNLI, SNLI, and MNLI) lie in the subsets where human agreement is low. However, because of the low human agreement, the model prediction will be nothing more than a random guess of the majority opinion. This raises an important concern that whether improving or comparing the performance

on this last missing part of the benchmarks is advisable or useful.

6 Discussion & Conclusion

While common practice in natural language evaluation compares the model prediction to the majority label, Section 5.4 questions the value of continuing such evaluation on current benchmarks as most of the unsolved examples are of low human agreement. To address this concern, we suggest NLP models be evaluated against the collective human opinion distribution rather than one opinion aggregated from a set of opinions, especially on tasks which take a descriptivist approach[10] to language and meaning, including NLI and common sense reasoning. This not only complements prior evaluations by helping researchers understand whether model performance on a specific data point is reliable based on its human agreement, but also makes it possible to evaluate models' ability to capture the whole picture of human opinions. Section 5.3 shows that such ability is missing from current models and potential room for improvement is huge.

It is also important to note that the level of human agreement is an intrinsic property of a data point. Section 5.4 demonstrates that such a property can be an indicator of the difficulty of the modeling. This hints at the connections between human agreements and uncertainty estimation or calibration (Guo et al., 2017) where machine learning models are required to produce the confidence value of their predictions, leading to important benefits in real-world applications.

In conclusion, we hope our data and analysis inspire future directions such as explicit modeling of collective human opinions; providing theoretical supports for the connection between human disagreement and the difficulty of acquiring language understanding in general; exploring potential usage of these human agreements; and studying the source of the human disagreements and its relations to different linguistic phenomena.

Acknowledgments

---

[10]Descriptivism is the prevailing trend in the current community compared to prescriptivism (Pavlick and Kwiatkowski, 2019).

References

Héctor Martínez Alonso, Barbara Plank, Arne Skjærholt, and Anders Søgaard. 2015. Learning to parse with iaa-weighted loss. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1361.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tongfei Chen, Zhengping Jiang, Keisuke Sakaguchi, and Benjamin Van Durme. 2019. Uncertain natural language inference. *ACL*.

Marie-Catherine De Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*, 38(2):301–333.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170.

Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 440–449. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562.

Lauri Karttunen, Stanley Peters, Annie Zaenen, and Cleo Condoravdi. 2014. The chameleon-like nature of evaluative adjectives. *Empirical issues in Syntax and Semantics*, 10:233–250.

Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning. 2006. Local textual inference: it's hard to circumscribe, but you know it when you see it–and nlp needs it. *Citeseer*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Nikita Nangia and Samuel R Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. In *ACL*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. *ACL*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pages 76–83.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of nli models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. In *EMNLP*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Lawrence Trask. 1999. *Key concepts in language and linguistics*. Psychology Press.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3-4):333–353.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

A   Annotation Interface

Figure 6 and 7 show the screenshots for NLI and αNLI collection, respectively.

B   Hyperparameters

For SNLI and MNLI, we used the same hyperparameters chosen by their original respective authors. For αNLI, we tuned batch size, learning rate and the number of epoch. For BERT, XLNet, and RoBERTa, we only searched parameters for large models and the base models use the same hyperparameters based on the results of the large ones. Table 8 shows the details.

C   Training Size and Trajectory

Figure 8 show the training trajectory and the changes of the accuracy and JSD of RoBERTa-large on four bins as the training data gradually increased in log space. The plots reveal that the accuracy of the models converges faster given fair amount of training data on bins with a high level of human agreements.

D   Label Statistics

Labeling statistics can be found in Table 7. It is worth noting that there is a shift of majority labels from neutral to entailment in MNLI-m. We assume the difference might be due to multi-genre nature of the MNLI dataset, and collecting more intuitive and concrete reasons for such an observation from a cognitive or linguistic perspective will be important future work.

E   Other Details

Our neural models are trained using a server with a Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz (10 cores) and 4 NVIDIA TITAN V GPUs. Table 6 shows the urls where we downloaded external resources.

Resource	URL
SNLI	https://nlp.stanford.edu/projects/snli
MNLI	https://cims.nyu.edu/~sbowman/multinli
αNLI	http://abductivecommonsense.xyz
ParlAI	https://parl.ai
Huggingface transformers	https://github.com/huggingface/transformers

Table 6: Links for external resources.

Collector:

Context: Under a blue sky with white clouds, a child reaches up to touch the propeller of a plane standing parked on a field of grass.

Statement: A child is reaching to touch the propeller of a plane.

Choose the correct category for the statement:

Given the context, the statement is **definitely correct**.

Given the context, the statement is **definitely incorrect**.

Given the context, the statement is **neither definitely correct nor definitely incorrect**.

The statement is **broken**.

Figure 6: Interface for NLI collection.

Collector:

O-Beginning: There was a scientist who wanted to collect bacteria samples.

Hypothesis-1: The scientist collected samples of the bacteria and tested them.

Hypothesis-2: He collected the bacteria and froze it.

O-Ending: The bacteria was non harmful.

Choose one of the hypothesises that is more likely to cause O-Beginning to turn into O-Ending.

Hypothesis-1 is more plausible than Hypothesis-2.

Hypothesis-2 is more plausible than Hypothesis-1.

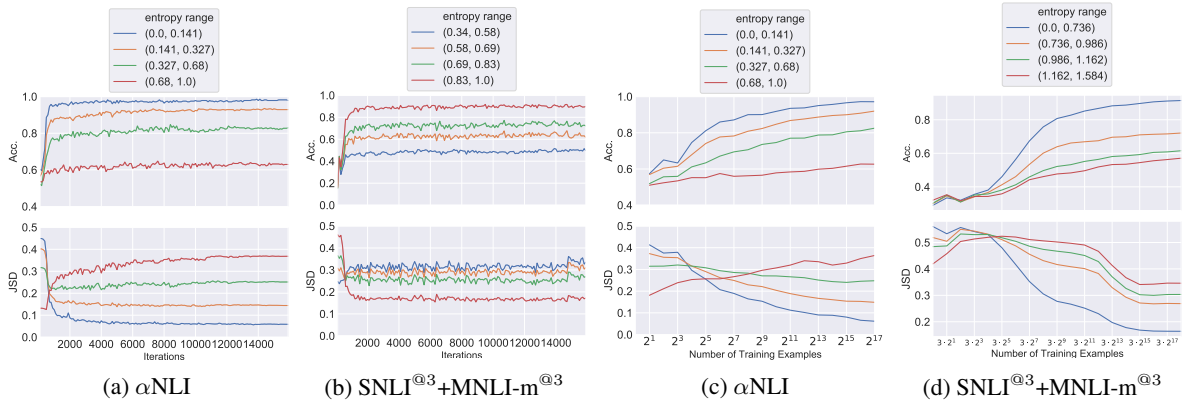Figure 7: Interface for αNLI collection.

(a) αNLI   (b) SNLI$^{@3}$+MNLI-m$^{@3}$ (c) αNLI   (d) SNLI$^{@3}$+MNLI-m$^{@3}$

Figure 8: Sub-figure 8a and 8b (the first two figures on the left) show the training trajectory of RoBERTa on αNLI and SNLI$^{@3}$+MNLI-m$^{@3}$. Sub-figure 8c and 8d (the two on the right) show the performance curves of RoBERTa on αNLI and SNLI$^{@3}$+MNLI-m$^{@3}$ as the training size increased.

Data	Label distribution (entailment / neutral / contradiction)			
	Old majority	New majority	Old raw count (5 per ex.)	New raw count (100 per ex.)
SNLI$^{@3}$	486 / 677 / 351	421$^{(-65)}$ / 813$^{(+136)}$ / 280$^{(-71)}$	2,470 / 3,420 / 1,673	45,113 / 76,063 / 30,185
MNLI-m$^{@3}$	513 / 721 / 365	741$^{(+228)}$ / 583$^{(-138)}$ / 275$^{(-90)}$	2,483 / 3,602 / 1,910	64,370 / 62,794 / 32,704

Table 7: NLI label distribution. 'Raw count' refers to the count of all individual labels. Superscript indicates the number of changes comparing to old majority labels.

Hyperparam {Search Range}	BERT	XLNet	RoBERTa	BART	ALBERT	DistilBert
Learning Rate {5e-5, 1e-5, 5e-6}	5e-5	5e-6	5e-6	5e-6	5e-6	5e-6
Batch Size {32, 64}	32	32	32	64	32	32
Weight Decay	0.0	0.0	0.0	0.01	0.0	0.0
Max Epochs {3, 4, 5}	5	5	3	5	5	5
Learning Rate Decay	Linear	Linear	Linear	Linear	Linear	Linear
Warmup ratio	0.1	0.1	0.1	0.1	0.1	0.1

Table 8: The best hyperparameters for finetuning models on αNLI.